DOCUMENT RESUME

ED 427 059 TM 029 440

AUTHOR	Evans, Victoria P.
TITLE	Strategies for Detecting Outliers in Regression Analysis: An Introductory Primer.
PUB DATE	1999-01-00
NOTE	40p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Identification; *Regression (Statistics)
IDENTIFIERS	*Outliers; *Residuals (Statistics)

ABSTRACT

Outliers are extreme data points that have the potential to influence statistical analyses. Outlier identification is important to researchers using regression analysis because outliers can influence the model used to such an extent that they seriously distort the conclusions drawn from the data. The effects of outliers on regression analysis are discussed, and examples of various detection methods are given. Most outlier detection methods involve the calculation of residuals. Given that the identification of a point as an outlier is not, in itself, grounds for exclusion, the questions that must be answered is when an outlying observation can be rejected legitimately. When individuals admit inattention during data collection, or acknowledge providing dishonest responses, the decision to delete outliers is straightforward. It is only troubling to delete them when the basis for the aberrance cannot be understood, and then the decision is the most difficult. Three appendixes contain a FORTRAN program to compute a type of detection matrix, input for that program, and output results for the example data. (Contains 4 tables, 6 figures, and 11 references.) (SLD)

Running head: OUTLIER IDENTIFICATION STRATEGIES

Strategies for Detecting Outliers in Regression Analysis: An Introductory Primer

Victoria P. Evans

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, 1999.

Abstract

Outliers are extreme data points that have the potential to influence statistical analyses. Outlier identification is important to researchers using regression analysis because outliers can influence the model used to such an extent that they seriously distort the conclusions drawn from the data. The present paper discusses the effects of outliers on regression analysis and offers examples of various detection methods.

In regression, outliers are data points with unusually large residuals (Anscombe, 1960). Data points that are outliers for some statistics (e.g., the mean) may not be outliers for other statistics (e.g., the correlation coefficient). For example, in the following score set, Alfred is an outlier on variable X (and on variable Y) as regards the mean, the standard deviation, and skewness and kurtosis, but not as regards the correlation coefficient.

	X	Y
Amanda	1	1
Jenny	3	2
Bob	5	3
Alfred	89	45

In statistical regression, as in all correlational analyses, outliers present particular challenges for researchers (Bacon, 1995). The following paper discusses the challenges of outliers in regression and presents examples of various outlier detection methods.

Regression Assumptions

For review, the assumptions underlying regression analysis (Hecht, 1991; Serdahl, 1996) are as follows:

1. a linear relationship must exist between variables;
2. the values of the dependent variable are distributed normally (follow the Gaussian distribution) for any values of the independent variable (Bump, 1991);
3. homoscedasticity; and

4. deviations from the least squares line of best fit are statistically independent.

Sources of Outliers

Outliers can exist in data sets for various reasons. One of the challenges of working with outliers is that the researcher is rarely sure of the reason for the outlying observations. Understanding the causes for outlying data points is important in the decision of whether to retain, eliminate, or recode the observations in question. The most common sources of outliers can be summarized as follows: population variability, measurement or recording errors (Anscombe, 1960; Iglewicz & Hoaglin, 1993), incorrect distributional assumptions, unaccounted for structure within the data (Iglewicz & Hoaglin, 1993), and execution error (Anscombe, 1960).

Population variability

If a distribution is distributed normally, then some variability must be present within the data. Under the normal distribution, a point can potentially exist anywhere within range of the distribution (Hecht, 1991, 1992). Simply because a data point is located a far distance from the mean does not necessarily imply that it is an errant observation nor that its existence calls into question the assumption of the general linear model (Hecht, 1991, 1992).

The extreme point may merely reflect natural variability within the population. Of course, aberrance due to this dynamic is more likely when sample size is small.

Measurement or Recording Errors

No measurement device nor researcher is completely infallible. At times, errors may be made in the measurement, or the recording or coding of the observation (Anscombe, 1960; Iglewicz & Hoaglin, 1993). Measurement apparatus may also be faulty (Anscombe, 1960). In these situations, if the researchers can be sure that the outlier was caused by measurement or recording errors, then they may legitimately choose to reject or recode the observation (Anscombe, 1960; Hecht, 1991). Iglewicz and Hoaglin (1993), however, advocate the recording of all outliers because if they reoccur in subsequent data collections, the reoccurrence may indicate the need to modify measurement or recording techniques.

Incorrect Distributional Assumptions

Outliers can appear in data sets if the distribution assumed for the analysis is incorrect (Iglewicz & Hoaglin, 1993). Points that are located large distances from the center may be more common in some distributions than in others. Therefore, assuming the correct distribution is important in research. Generally speaking, researchers should be aware of the distributional assumptions underlying regression analysis and deal with this matter before collecting data.

Structure Within the Data

Iglewicz and Hoaglin (1993) offer an example of data that are presumed to come from random daily samples but actually comes from a morning and an evening sample. In this case, the data may actually contain more structure than is being considered in the analysis. The data may need to be investigated more fully before deciding whether to retain, recode, or reject the outlying observations.

Execution Error

Anscombe (1960) pointed out that, as researchers, we do not always accomplish what we set out to accomplish. In other words, we may set out to measure one construct, but in actuality measure something slightly different.

Outlier Detection

By inspecting data for outliers, researchers can avoid making distorted conclusions about data and can make more robust estimates of parameters (Bacon, 1995). Iglewicz and Hoagin (1993) advocated the inspection of all data for outliers. Various outlier detection methods exist and will be discussed presently.

Residuals

Most outlier detection methods involve the calculation of residuals (i.e., Y - Yhat = e). In regression analysis, a squared residual (i.e., e^2) defines the amount of unexplained variability a given individual contributes to the total unexplained (within, error, residual) sum of squares, or the distance from the data point to the regression line on a scatter plot. One popular approach is to delete an observation if the magnitude of its residual exceeds the estimated population standard deviation multiplied by a designated constant (C). (Anscombe, 1960; Anscombe and Tukey, 1963; Hecht, 1991).

The value of the constant (C), is decided upon after careful consideration of the consequences of failure to reject erroneous observations versus mistaken rejection of good observations. To reject an observation, the magnitude of the residual of the observation must be large enough to exceed the product of C with the standard deviation (s). If C is large, then the largest residual will be less likely to exceed this product and the observation with the greatest residual will be less likely to be rejected. If C is small, however, the product of C with s will be smaller and the observation with the greatest residual will be more likely to be rejected. Researchers may choose a small value for C if they are greatly concerned with erroneous observations and wish to reject all of them. If, however, the data set includes no erroneous observations, there is no guarantee that no residuals will exceed the product of C with s. In such a case, good observations could potentially be rejected; and error variance of the parameter estimates would increase. The increase in error variance can be conceptualized as an insurance premium to protect against erroneous observations. Researchers must decide how much of a premium (increase in error variance) they are willing to pay to protect against erroneous observations. Usually a premium of 2 or 2.5% is considered to be an acceptable increase in error variance (Anscombe, 1960; Anscombe & Tukey, 1963). After the researcher decides upon the acceptable premium, C can be calculated. See Anscombe (1960) for more detailed information on the calculation of C.

The process of outlier rejection begins with the observation with the residual of the greatest magnitude and is recalculated after each deletion until no residuals remain with values greater than the magnitude of the constant times the standard deviation. To control for variation caused by the deleted outliers, the estimated population mean and standard deviation are recalculated each time the procedure is used (Anscombe, 1960; Anscombe & Tukey, 1963; Hecht, 1991).

Graphic Methods

Anscombe and Tukey (1963) encouraged beginning any analysis of residuals by looking at a scatterplot. Scatterplots usually show outliers as points located a far distance from the majority of the data points. Figure 1 shows a scatterplot for a set of data points to be used later in an investigation of outliers. Notice that one point in particular appears to be located a good distance from the perceived line of best fit for the other data points. The researcher in this case may suspect this data point of representing a grossly erroneous observation and choose to inspect it further.

Scatterplots of residuals against predictor variables can also help to detect outliers (Larsen & McCleary, 1972). These plots generally show outlier points located away from the center and can be inspected for model violations (Serdahl, 1996). Typically, residual plots are most helpful when the e scores are standardized so that the residual e scores are on the same scale as the y scores in their z score form (Serdahl, 1996). Figure 2 shows a scatterplot of standardized residuals against standardized predictor values. Twenty of the twenty-one points appear to fall close to a central line. The 21st data point, however, is located far away from the center of the others. Figure 2 was constructed from the same data used in Figure 1. The same observation is suspected of being an outlier in each plot.

<u>Hat Matrix</u>

Inspection of residual plots for large residuals can offer valuable information about outliers, but this method is not always completely effective. Some outliers exert enough influence on the regression line to make $yhat_i$ close to y_i. In such a case, the observation may perform as an outlier, but it does not have a large residual (Iglewicz & Hoaglin, 1993). The hat matrix can be helpful for detecting these types of cases.

The hat matrix maps y into yhat (Hoaglin & Welsch, 1978). Specifically, yhat = Hy, where $H = X(X^TX)^{-1}X^T$. The hat matrix is generally used to detect high leverage points, or points at which the magnitude of y has great influence on the fit. The term, h_{ij} of H, denotes the amount of leverage put forth on $yhat_i$ by y_j, or in other words, how changing y_j affects $yhat_i$. The diagonal of the hat matrix is composed of h_{ii} values. Each h_{ii} value represents the amount of leverage of an observed variable y_i on the corresponding latent variable $yhat_i$. The diagonal elements of the hat matrix express the influence of the observed variable y_i on the fit. The leverage of observation i is defined as h_{ii}: $h_{ii} =$

$$\frac{1}{N} + \frac{(x_i - xbar)^2}{E^n_{k=1}(x_k - xbar)^2}$$, where xbar is the mean

of the x variable.

High leverage points, or high values on the diagonal of the hat matrix, suggest that the corresponding observation may be an outlier. Typically, any h_{ii} greater than twice the number of predictors, or independent variables (IV), divided by the number of cases ($h_{ii} > 2IV/n$) can be considered to be a high leverage point (Hoaglin & Wesch, 1978). The following example, using the Draper and Stoneman data presented by Hoaglin and Welsch (1978), illustrates the utility of the hat matrix.

A FORTRAN program to compute the hat matrix was developed by Thompson (1998) and is presented in Appendix A. Two observed predictor variables and a constant were used as the independent variables for this example. Data for each independent variable and for the dependent variable were input into the FORTRAN program. The input data are presented in Appendix B. The first column of the Appendix lists the constant, the second and third columns list the observed values for the predictor scores, and the fourth column lists the scores on the dependent variable.

The hat matrix is computed on the basis of the formula: $yhat = X_{10x3}(X^T{}_{3x10}X_{10x3})^{-1}X^T{}_{3x10}y$. As seen in Appendix C, the input data matrix, X, has rank 10X3; the transpose of X has rank 3X10. Appendix C presents the product of X^T with X, the inverse of the product, and the product of X with the inverse matrix. A check was done to ensure that the product of the X^TX matrix with its inverse gave the identity matrix, indicating that the inverse of X^TX was truly the inverse.

The critical leverage value was determined by the 2IV/N rule, and was calculated to be 0.60. Notice the values (h_{ii}) on the diagonal of the hat matrix. Only the value for case 4 (0.6042) exceeds the critical value for leverage. This suggests that the observed score y for the fourth observation may influence the model fit. To determine the actual influence of case 4 on the model fit, standardized or studentized residuals may be computed for y_i when y_i is removed from the regression analysis.

Standardized Residuals

Residuals are usually expressed on a standard scale to facilitate interpretation (Hoaglin & Welsch, 1978; Iglewicz & Hoaglin, 1993). The formula for the adjustment is given by, $r_i/(s^2(1-h_i)^{1/2})$, where $e_i = r_i$ and s^2 is the residual mean square (Hoaglin & Welsch, 1978; Iglewicz & Hoaglin, 1993).

Studentized Residuals

Use of the studentized residual allows researchers to consider the extent to which an observation is an outlier by using statistical significance testing. The studentized residual involves the calculation of the residual of the data point in question when its influence has been removed from the data regression equation. The term $BETAhat_{(i)}$ is the least squares estimate of BETA on the data after observation i has been removed. The studentized residual is defined as,

$$r_i^* = \frac{y_i - x_i BETAhat_{(i)}}{s_{(i)}[1 + x_i(X^T_{(i)}X_{(i)})^{-1}x_i^T]^{1/2}}.$$

The resulting r_i* can then be used in a statistical significance test involving the t distribution to determine the statistical significance of the point's deviation from the remaining data.

Hoaglin and Welsch (1978) advocated the use of the hat matrix followed by an examination of studentized residuals. The hat matrix offers information about high leverage points, and the studentized residuals allow researchers to identify discrepant y values. Depending on the results of an examination of leverage points and residuals, researchers may choose to discard questionable data points, or, if the outlying data points are known to be accurate, the researchers may decide that the model does not adequately fit the data.

Influence of Outliers on Regression Analyses

Influence of y-Axis Outliers

Outliers on the dependent variable typically exert greater influence on the parameter estimates and R^2 value than do outliers on the independent variables (Hecht, 1991; Serdahl, 1996). Intuitively, this makes sense, as we consider that the outlying data point on y pulls the regression line towards itself in an effort to minimize error variance (Serdahl, 1996, p. 8). Hecht (1991) found that analysis of the standardized and studentized residuals were the most effective diagnostic methods for identifying outliers on the y axis.

Consider the fictitious data set presented in Table 1. Both the independent variable, x, and the dependent variable, y, were given equal means and standard deviations. There are no obvious outliers present in the data set. Figure 3 illustrates the output given by the SPSS computer package for regression statistics and outlier diagnostics. The output is presented exactly as it would appear when given by SPSS. Note that the R^2 value is an exceptionally high 92% and the BETA is .958.

Now consider the data in Tables 2 and 3. Table 2 presents a data set similar to the one given in Table 1, but the Table 2 data includes one extra case. The added case appears, upon visual inspection, to be an outlier on y but not x. To determine if the extra case truly is an outlying observation, we look to Figure 4. Notice under "Case Diagnostics" that case number 21 is listed with a standardized residual value of 4.133. Here, SPSS was asked to list only cases for which the standardized residual value exceeded 3.00, as reflected in the SPSS syntax file presented in Appendix A. This critical value is context-specific and may vary according to the study and researcher judgment. According to the criteria set by the present author, case 21 is a likely outlier for the given data set.

Notice that the R^2 value, .392, has suffered a 53% drop from the same value in Figure 1. The BETA value, .626, has also been reduced from the previous example. This change in BETA values indicates a considerable change in the regression equation used to predict values of the dependent variable once the single oulier on y has been dropped from the analysis.

Influence of x-Axis Outliers

Outliers on the x axis impact regression statistics, though to a smaller degree than do outliers on the y axis

(Hecht, 1991). Although outliers on x can, and do, influence the regression line, they usually have more effect on the variability of the x scores than they do on the relationship between the variables (Hecht, 1991).

Table 3 presents a data set with a suspiciously large value for x in the 21st case. The SPSS analysis presented in Figure 5, however, failed to report casewise diagnostics for this data because none of the observations yielded a standardized residual value greater than 3.00. Compare Table 2 with Table 3. Notice that the outlying observations would be identical except that value observed for y in Table 2 is the value observed for x in Table 3 and vice-versa. Interestingly, only the exceptional y value was considered to be an actual outlier. This discrepancy may exist because x is only considered for its impact on y. In other words, y is the variable of interest. The formula may be more sensitive to observations that are outliers on y than it is to observations that are outliers on x unless the extreme value of x shows a serious impact on y.

Influence of Both x-Axis and y-Axis Outliers

Table 4 presents data for a case in which both x and y appear, upon visual examination, to be outliers. The score for case 21 on y is identical to the score on the same variable in Table 2. In Table 2, case 21 was considered to be an outlier on y. In Figure 6, however, the case diagnostics section was omitted by SPSS indicating that no outliers were found for Table 4 data.

Notice the R^2 value in Figure 6 is 97% and the BETA coefficient is .983. Compared with Figure 1 for data with no obvious outliers, these values appear to be similar. The idea illustrated by this example is that even though data points may deviate from the mean, they may not necessarily impact the coefficient of determination or the regression equation. In this particular example, the data point in question is scaled in the same direction as the rest of the data. A graphical analysis would likely show little deviation of case 21 from the regression line involving only cases 1 through 20.

Identifying Damaging Outliers

Researchers need to recognize the distinction between outliers and damaging outliers. An observation that is identified as an outlier may or may not produce a damaging effect on the regression equation (Hecht, 1991). Rejecting or recoding data is rarely a desirable option because of the expense involved in data collection. Researchers should also be reluctant to reject data because they do not want to force data to conform to their preconceived hypotheses (Hecht, 1991).

Hecht (1991, 1992) asserted that too many researchers want to reject outlying observations simply because they are extreme points. Hecht contended that under the assumption of the Gaussian normal distribution, extreme data points have the potential to occur. To reject points simply because they are extreme is essentially to reject one of the assumptions upon which the regression analysis is based. If many extreme points occur in a data set, the assumption of the Gaussian distribution may need to be evaluated for violations. Hopefully, however, researchers would consider the distributional assumption before beginning the analysis.

Given that identification of a point as an outlier is not, in itself, grounds for exclusion, the question remains of when can one legitimately reject an outlying observation? Hecht (1991) advocated the rejection or recoding of an outlier when it is (a) due solely to measurement or recording errors or (b) when the outlier "hinders understanding by its inclusion in the model" (p. 22).

To determine the extent to which an outlier hinders understanding, researchers can compare two models, the first of which includes the extreme point in the construction of the model and the second of which does not.

The researchers must then decide whether or not they consider the difference in the two models to be meaningful from a contextual standpoint (Hecht, 1991; Hoaglin & Welsch, 1978).

Hoaglin and Welsch (1978, p. 20) suggested inspection of (a) the change in BETA weights from one model to the other or (b) the change in fit at the outlying point (x_i(BETAhat-BETA$_i$). In any case, the decision of whether or not to reject must be made from a contextual standpoint in light of all the data and distributional assumptions

However, outlier identification is not only a matter on blind dust-bowl empiricism. For example, when potential outliers are identified, when possible, it would be reasonable for the researcher to ask these persons whether they attended to the measurement tasks. The researcher might also explore reasons why these individuals behaved atypically; persons who responded honestly but unusually probably should be kept in the data set. When interviewing outlier candidates is not practical, sometimes researchers can nevertheless explore other information about these individuals to determine whether their behavior in retrospect seems reasonable.

However, when individuals admit inattention during data collection, or acknowledge providing dishonest responses, the decision to delete such outliers from further analysis is in this case straightforward. It is only troubling to delete outliers when the basis for the aberrance cannot be understood; that is when the decision of what to do with outliers is the most difficult.

References

Anscombe, F.J. (1960). Rejection of outliers. Technometrics, 2, 123-147.

Anscombe, F.J. & Tukey, J.W. (1963). The examination and analysis of residuals. Technometrics, 5, 141-160.

Bacon, D. (1995). A maximum likelihood approach to correlational outlier identification. Multivariate Behavioral Research, 30, 125-148.

Bump, W. (1991, January). The normal curve takes many forms: A review of skewness and kurtosis. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio. (ERIC Document Reproduction Service No. ED 342 790)

Hecht, J.B. (1991). Least squares regression and Schrodinger's cat: Perspectives on the analysis of regression residuals. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 333 020)

Hecht, J.B. (1992). Continuing perspectives on the analysis of Regression Residuals: Advocacy for the use of a 'Totter index'. Paper presented at the annual meeting of the American Educational Research Association, San

Francisco. (ERIC Document Reproduction Service No. ED 347 194)

Hoaglin, D.C. & Welsch, R.E. (1978). The hat matrix in regression and ANOVA. The American Statistician, 32, 17-22.

Iglewicz, B. & Hoaglin, D.C. (1993). How to detect and handle outliers. Milwaukee, WI: ASQC Quality Press.

Larsen, W.A. & McCleary, S.J. (1972). The use of partial residual plots in regression analysis. Technometrics, 14, 781-790.

Serdahl, E. (1996). An introduction to graphical analysis of residual scores and outlier detection in bivariate least squares regression analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 395 949)

Thompson, B. (1998). FORTRAN Program ("HATMAT.FOR") to Compute the "Hat Matrix" [Computer program]. College Station, TX: Author.

Table 1

Data for Regression Analysis With No Outliers

Case	x	y
1	5.00	5.00
2	5.00	5.50
3	6.00	5.50
4	6.50	6.50
5	6.50	6.25
6	6.25	6.50
7	6.50	7.00
8	7.00	6.50
9	7.00	7.50
10	7.50	7.00
11	7.50	7.00
12	7.50	8.00
13	7.50	7.75
14	7.75	7.50
15	8.00	8.50
16	9.00	8.50
17	9.00	9.00
18	9.50	10.00
19	9.50	9.00
20	10.00	10.00

Table 2

Data for Regression Analysis With Outlier on y

Case	x	y
1	5.00	5.00
2	5.00	5.50
3	6.00	5.50
4	6.50	6.50
5	6.50	6.25
6	6.25	6.50
7	6.50	7.00
8	7.00	6.50
9	7.00	7.50
10	7.50	7.00
11	7.50	7.00
12	7.50	8.00
13	7.50	7.75
14	7.75	7.50
15	8.00	8.50
16	9.00	8.50
17	9.00	9.00
18	9.50	10.00
19	9.50	9.00
20	10.00	10.00
21	7.50	15.00

Table 3

Data for Regression Analysis With Outlier on x

Case	x	y
1	5.00	5.00
2	5.00	5.50
3	6.00	5.50
4	6.50	6.50
5	6.50	6.25
6	6.25	6.50
7	6.50	7.00
8	7.00	6.50
9	7.00	7.50
10	7.50	7.00
11	7.50	7.00
12	7.50	8.00
13	7.50	7.75
14	7.75	7.50
15	8.00	8.50
16	9.00	8.50
17	9.00	9.00
18	9.50	10.00
19	9.50	9.00
20	10.00	10.00
21	15.00	7.50

Table 4

Data for Regression Analysis With Outliers on x and y

Case	x	y
1	5.00	5.00
2	5.00	5.50
3	6.00	5.50
4	6.50	6.50
5	6.50	6.25
6	6.25	6.50
7	6.50	7.00
8	7.00	6.50
9	7.00	7.50
10	7.50	7.00
11	7.50	7.00
12	7.50	8.00
13	7.50	7.75
14	7.75	7.50
15	8.00	8.50
16	9.00	8.50
17	9.00	9.00
18	9.50	10.00
19	9.50	9.00
20	10.00	10.00
21	15.00	15.00

Figure 1

Scatterplot for Table 3 Data

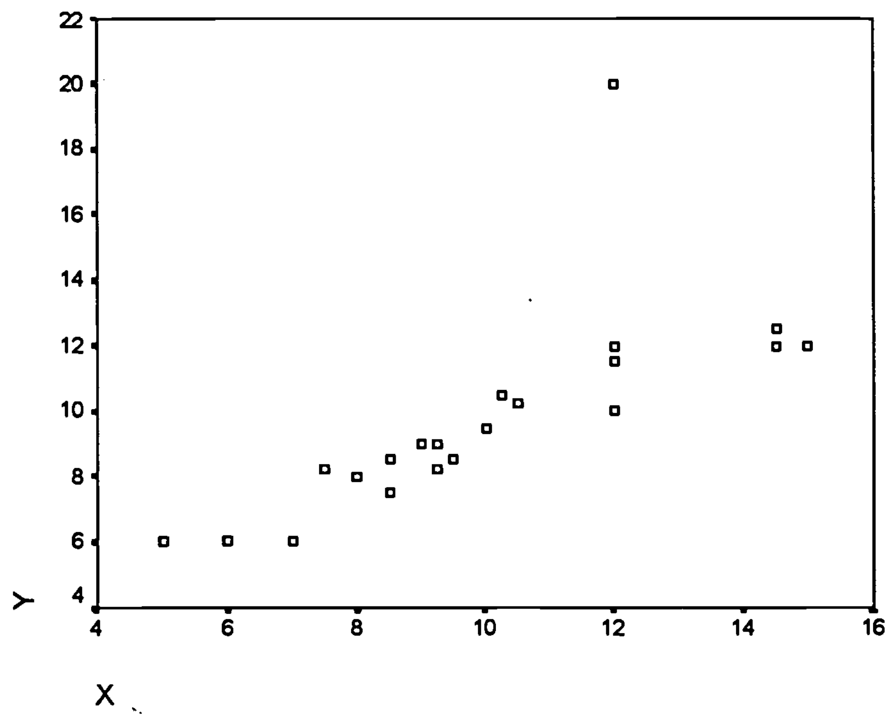

Figure 2

Scatterplot of Standardized Residuals Against Standardized Predictor Values

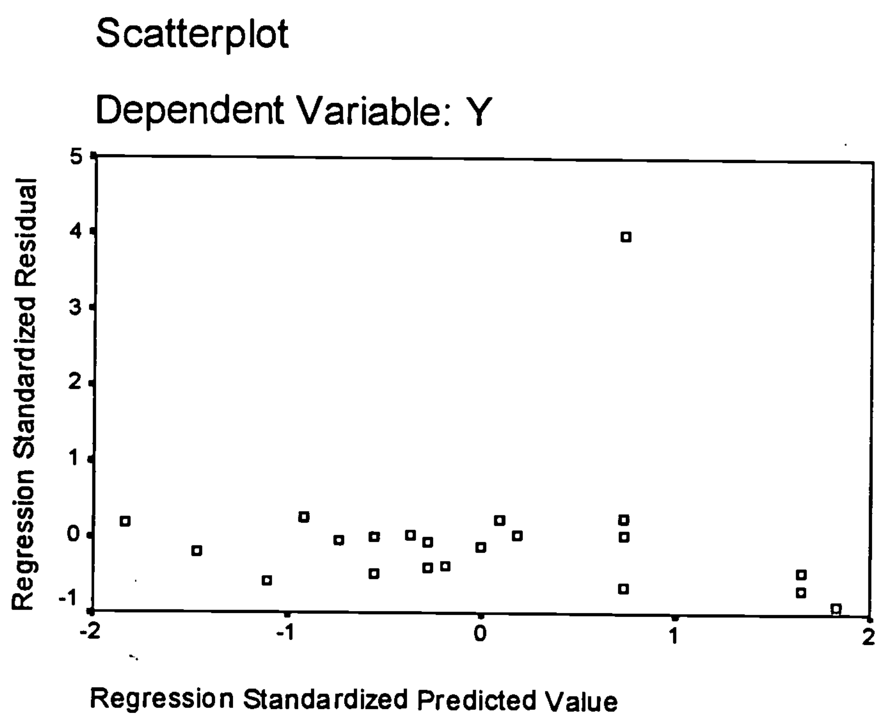

Figure 3

SPSS Output for Table 1 Data With No Outliers

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.958[a]	.918	.913	.4204

a. Predictors: (Constant), X

b. Dependent Variable: Y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35.581	1	35.581	201.281	.00
	Residual	3.182	18	.177		
	Total	38.763	19			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.311	.510		.610	.54
	X	.958	.068	.958	14.187	.00

a. Dependent Variable: Y

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5.1017	9.8921	7.4250	1.3685	2
Residual	-.5597	.5870	-1.78E-16	.4092	2
Std. Predicted Value	-1.698	1.803	.000	1.000	2
Std. Residual	-1.331	1.396	.000	.973	2

a. Dependent Variable: Y

Figure 4

SPSS Output for Table 2 Data with Outlier on y

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.958[a]	.918	.913	.4204

a. Predictors: (Constant), X

b. Dependent Variable: Y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35.581	1	35.581	201.281	.00
	Residual	3.182	18	.177		
	Total	38.763	19			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.311	.510		.610	.54
	X	.958	.068	.958	14.187	.00

a. Dependent Variable: Y

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5.1017	9.8921	7.4250	1.3685	2
Residual	-.5597	.5870	-1.78E-16	.4092	2
Std. Predicted Value	-1.698	1.803	.000	1.000	2
Std. Residual	-1.331	1.396	.000	.973	2

a. Dependent Variable: Y

Figure 5

SPSS Output for Table 3 Data With Outlier on x

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.958[a]	.918	.913	.4204

a. Predictors: (Constant), X

b. Dependent Variable: Y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35.581	1	35.581	201.281	.00
	Residual	3.182	18	.177		
	Total	38.763	19			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.311	.510		.610	.54
	X	.958	.068	.958	14.187	.00

a. Dependent Variable: Y

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5.1017	9.8921	7.4250	1.3685	2
Residual	-.5597	.5870	-1.78E-16	.4092	2
Std. Predicted Value	-1.698	1.803	.000	1.000	2
Std. Residual	-1.331	1.396	.000	.973	2

a. Dependent Variable: Y

Figure 6

SPSS Output for Table 4 Data With Outliers on Both x and y

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.958[a]	.918	.913	.4204

a. Predictors: (Constant), X

b. Dependent Variable: Y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35.581	1	35.581	201.281	.00
	Residual	3.182	18	.177		
	Total	38.763	19			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.311	.510		.610	.54
	X	.958	.068	.958	14.187	.00

a. Dependent Variable: Y

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5.1017	9.8921	7.4250	1.3685	2
Residual	-.5597	.5870	-1.78E-16	.4092	2
Std. Predicted Value	-1.698	1.803	.000	1.000	2
Std. Residual	-1.331	1.396	.000	.973	2

a. Dependent Variable: Y

APPENDIX A
FORTRAN Program ("HATMAT.FOR") to Compute the "Hat" Matrix

```
C234567   HATMAT.FOR 10/29/98
C  Hoaglin, D.C., & Welsch, R.E. (1978). The Hat Matrix in
C     regression and ANOVA. _The American Statistician_,
C     _32_(1), 17-22.
C
      INTEGER XR,XC,XTR,XTC,XTIR,XTIC
      REAL TIT(20),X(500,30),XT(30,500),STOR(30,30),
     *XTINV(30,30),STOR2(500,30),PROD(500,500),Y(500),
     *IF(20)
      IN=99
      IO=27
      XR=500
      XC=30
      XTR=XC
      XTC=XR
      XTIR=XC
      XTIC=XTIR
C
      READ(IN,1)TIT,N,IV
    1 FORMAT(20A4/2I5)
      WRITE(IO,2)TIT,N,IV
    2 FORMAT(/' JOB TITLE: ',20A4/' N= ',I5/
     *' N OF IVs= ',I5)
      READ(IN,3)IF
    3 FORMAT(20A4)
      WRITE(IO,4)IF
    4 FORMAT(' READ FORMAT: ',20A4//' Input Data:')
      DO 5 I=1,N
      READ(IN,IF)(X(I,J),J=1,IV)
    5 WRITE(IO,6)I,(X(I,J),J=1,IV)
    6 FORMAT(1X,I5,1X,10F8.3/9(7X,10F8.3/))
C
C TRANSPOSE X
      DO 7 I=1,N
        DO 8 J=1,IV
        XT(J,I)=X(I,J)
    8   CONTINUE
    7 CONTINUE
      WRITE(IO,9)
    9 FORMAT(//' Transpose of X:')
      DO 10 J=1,IV
   10 WRITE(IO,6)J,(XT(J,I),I=1,10)
C
C XT times X
      CALL MRRRR(IV,N,XT,XTR, N,IV,X,XR, IV,IV,STOR,XTR)
      WRITE(IO,11)
```

```
   11 FORMAT(//' XT times X:')
      DO 12 I=1,IV
   12 WRITE(IO,6)I,(STOR(I,J),J=1,IV)
C
C Invert XT times X
      CALL LINRG(IV,STOR,XTR, XTINV,XTIR)
      WRITE(IO,13)
   13 FORMAT(//' Inverse of XT times X:')
      DO 14 I=1,IV
   14 WRITE(IO,6)I,(XTINV(I,J),J=1,IV)
C
C Check inverse
      CALL MRRRR(IV,IV,STOR,XTR, IV,IV,XTINV,XTIR,
     *           IV,IV,STOR2,XR)
      WRITE(IO,15)
   15 FORMAT(//' Check if Inverse yields I matrix:')
      DO 16 I=1,IV
   16 WRITE(IO,6)I,(STOR2(I,J),J=1,IV)
C
C Multiply X times XTINV
      CALL MRRRR(N,IV,X,XR, IV,IV,XTINV,XTIR, N,IV,STOR2,XR)
      WRITE (IO,17)
   17 FORMAT(//' X times XTINV:')
      L=20
      IF(N.LT.L)L=N
      DO 18 I=1,L
   18 WRITE(IO,6)I,(STOR2(I,J),J=1,IV)
C
C Compute PROD matrix
      CALL MRRRR(N,IV,STOR2,XR, IV,N,XT,XTR, N,N,PROD,XR)
      WRITE(IO,19)
   19 FORMAT(//' The HAT matrix result:')
      DO 20 I=1,N
      Y(I)=PROD(I,I)
   20 WRITE(IO,21)I,(PROD(I,J),J=1,N)
   21 FORMAT(1X,I5,1X,10F8.4/99(7X,10F8.4/))
C
C Compute 'rule of thumb' critical value
      CRIT=(2.*FLOAT(IV))/FLOAT(N)
      WRITE(IO,22)CRIT
   22 FORMAT(//' The rough critical value for leverage'/
     *' ((2 x IV) / N) = ',F8.5//)
      WRITE(IO,23)
   23 FORMAT(//'The diagonal leverage values:')
      DO 24 I=1,N
      IF(Y(I).LE.CRIT)WRITE(IO,25)I,Y(I)
   25 FORMAT(1X,I5,1X,F8.4)
      IF(Y(I).GT.CRIT)WRITE(IO,26)I,Y(I)
   26 FORMAT(1X,I5,1X,F8.4,' ****')
   24 CONTINUE
      WRITE(IO,27)
```

```
   27  FORMAT(//' Asterisks designate leverage above "rule of
thumb."'/)
C
C
 9999 STOP
      END
```

APPENDIX B
Input into the HATMAT.FOR Program
Using the Hoaglin and Welsch (1978) Example Data

```
Hoaglin & Welsch (1978). _Am Stat_, _32_(1), 17-22.
   10    3
 (F6.3,F6.3,F5.1,F6.2)
10.302 0.499 11.1 11.14
10.302 0.558  8.9 12.74
10.302 0.604  8.8 13.13
10.302 0.441  8.9 11.51
10.302 0.550  8.8 12.38
10.302 0.528  9.9 12.60
10.302 0.418 10.7 11.13
10.302 0.480 10.5 11.70
10.302 0.406 10.5 11.02
10.302 0.467 10.7 11.41
```

Note. 10.302 is the additive constant for these data to predict "strength" using the "specific gravity" and "moisture content" variables in the article presented by Hoaglin and Welsch (1978).

APPENDIX C
Output Results for the Example Data

```
JOB TITLE: Hoaglin & Welsch (1978). _Am Stat_, _32_(1), 17-22.
N=    10
N OF IVs=      3
READ FORMAT:  (F6.3,F6.3,F5.1,F6.2)

Input Data:
    1   10.302   0.499  11.100
    2   10.302   0.558   8.900
    3   10.302   0.604   8.800
    4   10.302   0.441   8.900
    5   10.302   0.550   8.800
    6   10.302   0.528   9.900
    7   10.302   0.418  10.700
    8   10.302   0.480  10.500
    9   10.302   0.406  10.500
   10   10.302   0.467  10.700


Transpose of X:
    1   10.302  10.302  10.302  10.302  10.302  10.302  10.302
        10.302  10.302  10.302

    2    0.499   0.558   0.604   0.441   0.550   0.528   0.418
         0.480   0.406   0.467

    3   11.100   8.900   8.800   8.900   8.800   9.900  10.700
        10.500  10.500  10.700


XT times X:
    1  1061.312    51.005 1017.837
    2    51.005     2.489   48.585
    3  1017.837    48.585  984.000


Inverse of XT times X:
    1     0.447  -3.714  -0.279
    2    -3.714  41.986   1.769
    3    -0.279   1.769   0.202


Check if Inverse yields I matrix:
    1     1.000   0.003   0.000
    2     0.000   1.000   0.000
    3     0.000   0.001   1.000
```

X times XTINV:

1	-0.345	2.321	0.253
2	0.049	0.908	-0.086
3	-0.094	2.662	-0.025
4	0.484	-4.005	-0.293
5	0.107	0.395	-0.121
6	-0.118	1.417	0.062
7	0.068	-1.787	0.029
8	-0.107	0.463	0.098
9	0.168	-2.644	-0.032
10	-0.114	0.270	0.116

The HAT matrix result:

1	0.4178	-0.0020	0.0795	-0.2736	-0.0459	0.1814	0.1285	0.2219	0.0501	0.2423
2	-0.0020	0.2419	0.2923	0.1357	0.2433	0.1281	-0.0409	0.0327	-0.0345	0.0036
3	0.0795	0.2923	0.4173	-0.0192	0.2735	0.1871	-0.1260	0.0441	-0.1529	0.0044
4	-0.2736	0.1357	-0.0192	0.6042	0.1970	-0.0376	0.1681	-0.0215	0.2749	-0.0281
5	-0.0459	0.2433	0.2735	0.1970	0.2522	0.1106	-0.0295	0.0191	-0.0101	-0.0102
6	0.1814	0.1281	0.1871	-0.0376	0.1106	0.1479	0.0418	0.1172	0.0123	0.1112
7	0.1285	-0.0409	-0.1260	0.1681	-0.0295	0.0418	0.2616	0.1450	0.2773	0.1741
8	0.2219	0.0327	0.0441	-0.0215	0.0191	0.1172	0.1450	0.1540	0.1198	0.1677
9	0.0501	-0.0345	-0.1529	0.2749	-0.0101	0.0123	0.2773	0.1198	0.3155	0.1477
10	0.2423	0.0036	0.0044	-0.0281	-0.0102	0.1112	0.1741	0.1677	0.1477	0.1873

The rough critical value for leverage
((2 x IV) / N) = 0.60000

The diagonal leverage values:

1	0.4178	
2	0.2419	
3	0.4173	
4	0.6042	****
5	0.2522	
6	0.1479	
7	0.2616	
8	0.1540	
9	0.3155	
10	0.1873	

Asterisks designate leverage above "rule of thumb."

TM029440

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE
(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: STRATEGIES FOR DETECTING OUTLIERS IN REGRESSION ANALYSIS: AN INTRODUCTORY PRIMER

Author(s):

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

[X] Sample sticker to be affixed to document

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
VICTORIA P. EVANS
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document [ ]

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: X Victoria P. Evans	Position: RES ASSOCIATE
Printed Name: VICTORIA P. EVANS	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1831
	Date: 12/3/98